Visualization and analysis of large data collections: a case study applied to confocal microscopy data

W. de Leeuw[*]
Center for Mathematics and Computer Science CWI

P. J. Verschure[†]
Swammerdam Institute for Life Sciences SILS

R. van Liere[‡]
Center for Mathematics and Computer Science CWI

Abstract

In this paper we propose an approach in which interactive visualization and analysis are combined with batch tools for the processing of large data collections. Large and heterogeneous data collections are difficult to analyze and pose specific problems to interactive visualization. Application of the traditional interactive processing and visualization approaches as well as batch processing encounter considerable drawbacks for such large and heterogeneous data collections due to the amount and type of data. Computing resources are not sufficient for interactive exploration of the data and automated analysis has the disadvantage that the user has only limited control and feedback on the analysis process. In our approach, an analysis procedure with features and attributes of interest for the analysis is defined interactively. This procedure is used for offline processing of large collections of data sets. The results of the batch process along with "visual summaries" are used for further analysis. Visualization is not only used for the presentation of the result, but also as a tool to monitor the validity and quality of the operations performed during the batch process. Operations such as feature extraction and attribute calculation of the collected data sets are validated by visual inspection. This approach is illustrated by an extensive case study, in which a collection of confocal microscopy data sets is analyzed.

CR Categories: I.3.3 [Computer Graphics]: Picture/Image Generation; I.3.7 [Computer Graphics]: Methodology and Techniques

Keywords: biomedical visualization, features in volume data sets, large data set visualization.

1 Introduction

Interactive visualization is a valuable tool for data analysis. However interactive visualization tools are limited with respect to the amount of data that can be handled. Much work has been done to speed up visualization algorithms of large volume data sets (e.g., iso-surface extraction from a high resolution simulation). However, in many cases the interest of a researcher is not a single data set but a collection of individual data sets to be analyzed simultaneously. For example, for a simulation parameter study where the simulation produces a data set for each parameter setting, both the interaction with the data as a collection as well as the interaction with individual data is needed.

Data generated for research purposes is often used to test hypotheses. In this case the goal of data acquisition and the following analysis is based on confirmation or rejection of the hypothesis. To test the hypotheses, the input data has to be interpreted in the context of the experiment. This is done by selection and quantification of certain aspects of the data. In a number of steps the input data is transformed to an answer to the research question.

In many research experiments, collections of data have to be analyzed. The number of performed measurements for a particular experiment depends on the variability of the studied phenomenon and the number of aspects that are studied in the experiment. The total amount of data can become very large when the raw measurements are image or volume data. The analysis of such data poses specific demands on interactive visualization. If the collection is large it is very time consuming to process each set separately in an interactive session. The ability to inspect the analysis process by checking the intermediate results and by tracing back the results to the raw data is required for analysis of large data collections.

When using digital microscopy, often large data collections are produced. Digital microscopy such as confocal laser scanning microscopy (CLSM) allows biologists and biomedical end-users to obtain high resolution 3D data sets of biological objects, such as cells and tissues. These data are used to elucidate molecular mechanisms of cell control.

In most experiments involving CLSM, conclusions can not be drawn from a single data set due to biological diversity. To compensate for the biological diversity, a significant number of images have to be analyzed. Moreover, in biological experiments not all parameters can be controlled. Repetition of an experiment leads to a large variation in the microscopic images. Conclusions are based on results of the analysis of this collection of data sets. The total amount of data to be analyzed in a typical experiment is in the order of tens or even hundreds of gigabytes. For quantitative analysis and further interpretation of microscopic data, tools are used to extract relevant measures from the data. Visualization is crucial here, not only for the presentation of the final result, but also as a tool to inspect the analysis process used to obtain the results.

In a typical volume analysis system a single volume data set is processed at a time. Processing of each set separately is time consuming. However, as each data set might have peculiarities not foreseen during the preparation of the calculation, inspection of the process is required. Standard off line methods are not suitable because with these methods only the final results can be accessed and inspection of intermediate stages is difficult if not impossible. Thus on the one hand more monitoring and interaction of the analysis process is needed than off line batch processing offers. On the other hand, the amount of data is too large for analysis using fully interactive visualization tools. In the present paper, we describe tools and methods to combine these processing types. These tools were implemented in a system for the efficient visual analysis of potentially very large collections of CLSM data sets.

In the next section our developed system is put in the context of related studies. Section 3 presents the challenges presented by the analysis of large data collections. Section 4 gives an overview of problems and visualization requirements specific to the analysis

[*] e-mail: wimc@cwi.nl
[†] e-mail: pj.verschure@science.uva.nl
[‡] e-mail: robertl@cwi.nl

of CLSM data and in particular the visualization and interaction tools for the analysis of collections of volume data sets. Section 5 discusses a case study which illustrates the discussed concepts in a biological experiment. In the last section the conclusions of the present issue are drawn.

2 Related work

Several researchers worked on problems related to visualization of CLSM data. Their work concentrates on the presentation of a volume through volume rendering techniques and the aspects specific to CLSM data. For instance, Kaufman et al. [20] present BioCube, a system specifically tailored towards the display of biological volumetric data. Aspects such as the signal to noise ratio of the biological data, and the fact that the viewer usually has little a priori knowledge about the data are addressed by Kaufman et al. by using shading algorithms in which the user can control the trade-off between noise suppression and loss of detail. In the paper by Sakas et al. [23], problems specific to the presentation of volume data acquired by CLSM such as low contrast and unequal resolution in the plane and depth direction, are addressed. Median filtering of data is used to suppress noise while maintaining sharpness of contours to improve rendering quality. Transfer function design for, among other types, biological data is the main focus of the work Fang et al. [12, 5]. By using an image based transfer function model that integrates 3D image processing tools into the volume visualization pipeline the user can adjust the transfer function in and intuitively clear fashion. Segmentation and quantification of biological data has been an active research topic in recent years. Techniques such as level sets [29] improved the possibilities to extract relevant structures from noisy volume data. Fernadez-Gonzales et el. [13] and Knowles at el. [21] extensively studied the segmentation and quantification of cellular structures from volume data acquired by CLSM data for diagnostic purposes.

A number of systems and toolkits exist for the analysis and visualization of volume data. ITK [2] is an open-source software toolkit for performing registration and segmentation on volume data acquired for instrumentation such as CT or MRI scanners. Amira [11] originates from the Department for Scientific Visualization of the Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB). It supports segmentation tools, reconstruction algorithms to create polygonal models from segmented objects and visualization methods for the processing of 3D data sets from medicine, biology, physics or engineering. The system contains tools and support for interactive visualization and segmentation of multiple CLSM data sets. Beside the described systems there are many others (commercial) 3D volume visualization systems which are tailored in various degrees to the visualization of biological or confocal data such as Volview [3], Imaris [1], and VoxBlast [4] are

The idea of giving an overview of the process which lead to the final (visualization) result can also be found in work by Jankun-Kelly and Ma [18, 19]. The combination of computation intensive processing and interactive visualization is the main goal of computational steering [14]. In computational steering computation and visualization occur simultaneously and the user can interact with the calculation for example by changing simulation parameters.

Our approach to visualize CLSM data differs from the above mentioned approaches in that we specifically focus on dealing with large collections of data sets. Our goal is to improve the visualization of computation intensive analysis with interactive tools.

3 Batch and interactive processing

Batch processing and interactive analysis are two approaches for analysis of large data collections. These approaches are illustrated in figure 1. During interactive analysis the data along with the analysis tools and visualization tools are available to the user. The user can control the various processing parameters in the processing steps directly and the user can observe the effect of changes that are introduced. In batch processing a setup is prepared beforehand, i.e., the operations to apply and the processing parameters are defined. Moreover, this set of operations is applied to all data and the results of the calculation are stored and can be inspected afterwards.

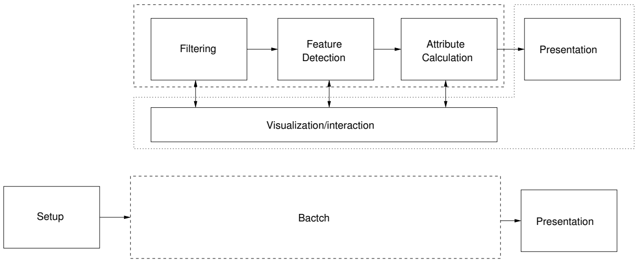


Figure 1: Comparison between interactive (top) analysis and batch processing (below).

The concept of a “visual summary” is used as a way to integrate batch and interactive processing. A visual summary is an image or set of images which gives the user the ability to inspect the batch process without the need to store all data generated during processing. They are produced during the batch processing and used during analysis of the batch output to present an overview of what happened during processing of the data. They can be seen as a layer of information between the information extracted during the batch process and the raw data which is used for the analysis. This information can be used during analysis of the batch output to get a sense of the data on which a particular data point is based, without the need of accessing the original data or re-executing the applied process.

Interactive tools can be utilized to setup the parameters and filters to be applied to the data in a batch process. Interactive setup of the batch process using a single or a sample data set can help to speed up the creation of a setup stage. Furthermore it can reduce repetition of processing due to incorrectly guessed parameters.

4 Confocal Laser Scanning Microscopy

Microscopy is an essential tool in biological and biomedical research. Digital microscopy such as CLSM allows biologists and biomedical end-users to obtain high resolution 3D data sets of biological objects, such as cells and tissues. Moreover, new luminescent probing techniques to visualize intra-cellular components and processes in living cells are becoming available. By using probes with fluorescent labels that have different spectral properties, different biological structures can be visualized in a single specimen. These developments are used in experiments to understand the molecular mechanisms of cell control.

In CLSM data is obtained by scanning a fluorescently labeled specimen using a laser beam. The laser beam, steered by a deflection mechanism and focused by an objective, scans the specimen in a regular fashion. The reflected light is captured by a photo detector via a beam splitter. A pinhole in front of the detector blocks the light from outside the focal plane. In this way a volumetric data set from the data is constructed. The data is scanned with a voxel size in the order of 50 nm in the x/y direction and 200 nm in the z direction. A typical result consists of 3 fluorescently labeled structures captured in 3 independent CLSM channels with a size of 512x512x60 voxels with 12 bits of intensity information, generating approximately 60

Mb of data. The resolution of the data will increase further in the near future since imaging techniques are still improving.

In biological experiments not all parameters can be controlled because cells are living entities. For example, variations in the life cycle of a cell cause many variations in the appearance of biological structures causing unavoidable variations in data sets representing the same structure. Furthermore, obtained data sets contain a significant amount of noise. In principle, conclusions can not be drawn from a single data set but have to be based on a collection of data sets. The total amount of data to be analyzed in an experiment is in the order of several tens up to hundreds of gigabytes.

4.1 Analysis

The analysis process of microscopy data can be divided into four stages:

Restoration

Image restoration addresses the effect of the instrumentation on the data. This includes corrections for distortions introduced by the microscope such as light from out of focus planes or chromatic aberration. Deconvolution algorithms and shift estimation algorithms are available to correct for these distortions.

Segmentation

In this stage the features relevant to the experiment are extracted from the restored data. For an experiment a specific type of feature of the data is of interest. Image processing tools are used to extract these features. Filters such as Gaussian filters and band-pass filters are used to filter certain frequencies from the data. Segmentation of the filtered data is accomplished by application of a boolean expression on the voxel values which decides of a voxel is part of the feature. Objects in the data are detected by checking for connectivity of voxels.

Attribute calculation

Attributes are characteristics of the data set or of features detected in the data. For instance, attributes are the average signal value of a data set, the volume of a feature, or the ratio between surface and volume of a feature.

Presentation

The results of the analysis have to be presented to the experimenter or a wider audience. In external presentations the outcome of an experiment is communicated mostly using easy to understand graphs of simple data types. For the purpose of inspection and of, if possible, interaction with the analysis process, more complicated data types and relations between the processed parameters have to be presented to the user. The results of the analysis have a variety of data types e.g., the processed volumes, the features, and the attributes each have specific data types.

4.2 The Argos system

The possibility to combine batch processing with interactive methods using visual summaries was implemented in a system named Argos. It is a system for the analysis of collections of volume data sets, with a focus on problems specific to the analysis of CLSM data sets. The input into the system consists of a collection of volume data sets. Volume data sets are processed in the Argos system by an interactively constructed feature extraction and attribute calculation pipeline consisting of a series of operators. The operators and parameters to be used depend on the research question. The user can select and connect operators needed for an experiment from a library of operators. The operators have parameters which can be adjusted to suit a particular purpose. Examples of operators are a band-pass filter, spot detection using a fixed threshold and calculation of the size of a feature.

The results of a selection operation which consists of a number of disjoint regions in the data volume are referred to as features. Each feature can be addressed by operations which can be used to calculate *feature attributes*.

The pipeline of operators is used to process data sets in the collection under control of the user. The analysis can be applied to a single data set keeping intermediate results or it can be applied to all data sets. The results of the analysis are presented in views depending on the type of operator results. The volume and shape of features are presented in a volume viewer, whereas attribute data are viewed in a graphing environment. A screen shot showing a typical interaction session is shown in figure 2.

The system handles a collection consisting of an arbitrary number of individual data sets. Each data set consists of a number of CLSM channels that consist of structures that are imaged by excitation of fluorescent labels with different spectral properties. For each channel attributes can be calculated. Each data set has a number of feature collections calculated by a feature filter. The feature collection is composed of the individual features. Particular attributes of the extracted features (e.g., the average intensity gradient magnitude, roundness of the area, the number local maximum values, the size of the region around the local maximum) can be calculated.

4.3 Interaction and processing tools

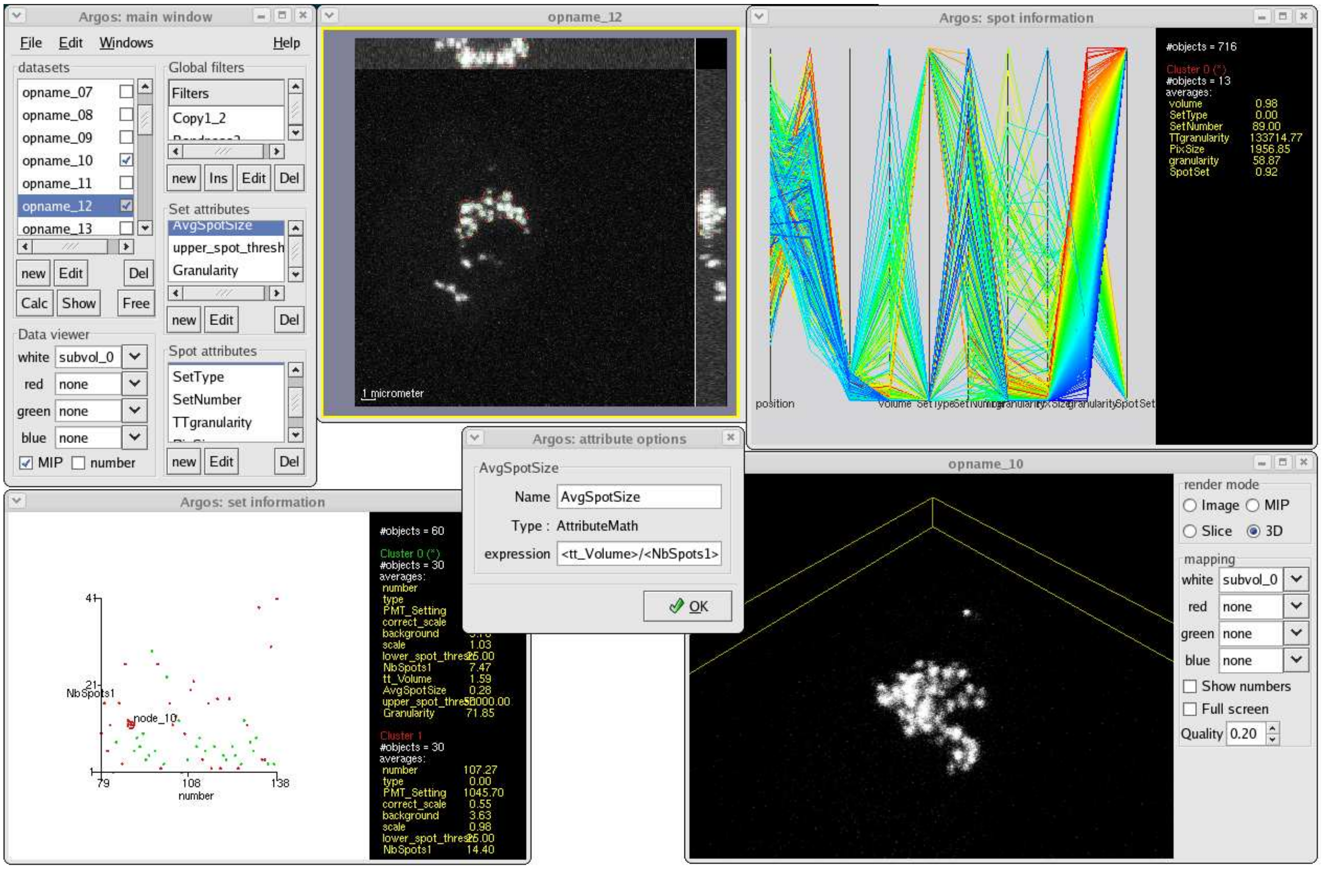
The amount of time needed to process each data set from a collection individually is such that analysis of a collection set by set is not practical. The processing time for a single data set is typically in the order of minutes, depending on the machine used and the attributes needed for the particular experiment. A collection of a hundred of such sets takes several hours. Running the calculation without user intervention in batch mode is preferred. During this process, the features characterized by their attributes are stored.

The user has to decide which filters and segmentation algorithms have to be applied, which attributes have to be calculated, and what parameter settings for the operators are used, for the batch processing of a collection of data sets. In Argos the user can generate a set up interactively, and test it on a single data set or on a particular data selection from the whole collection. Processing of the data using a collection of data sets and a prepared set up can be done in the interactive environment. This option gives the user the possibility to monitor the calculation as it is performed.

After processing the user can only inspect data saved during the batch process. Storing all intermediate data would increase the storage requirements an order of magnitude compared to the raw data which is already considerable large and is therefore not feasible. Feedback on the processing of individual data sets is needed to be able to judge the quality of the feature extraction and attribute calculation for a particular data set.

The visual summaries are generated during batch processing. The visual summaries are treated as operations like other operations which can be applied to the data, and they are calculated and stored just like the other operations. The visual summary operation produces an image of the data set using a restricted set of the same mapping parameters used for visualization during interactive processing. When the results of the batch process are inspected, this image can be used to present information with respect to the batch process. In case of a complicated filtering and extraction processes more visual summary operations can be defined to highlight the different aspects of the batch process. The additional image information such as the physical dimensions of the visualized area is stored during the batch process. This information is used during inspection to show the locations of features.

The images of the visual summaries are generated in the same way as in the interactive setting. In the interactive mode the user can make an arbitrary mapping of data channels or feature sets to color. This makes inspection easy as they are presented in the same way as during interactive visualization, with the only difference being that the interaction possibilities are limited. The mapping of



AvgSpotSize

Name AvgSpotSize

Type : AttributeMath

expression <tt_Volume>/<NbSpots1>

Quality 0.20

```
#objects = 716

Cluster 0 (*)
#objects = 13
averages:
 volume          0.98
 SetType         0.00
 SetNumber       89.00
 TTgranularity   133714.77
 PixSize         1956.85
 granularity     58.87
 SpotSet         0.92
```

```
#objects = 60

Cluster 0 (*)
#objects = 30
averages:
 number
 type
 PMT_Setting
 correct_scale
 background
 scale             1.03
 lower_spot_thres25.00
 NbSpots1          7.47
 tt_Volume         1.59
 AvgSpotSize       0.28
 upper_spot_thre50000.00
 Granularity       71.85

Cluster 1
#objects = 30
averages:
 number            107.27
 type              0.00
 PMT_Setting       1045.70
 correct_scale     0.55
 background        3.63
 scale             0.98
 lower_spot_thres25.00
 NbSpots1          14.40
```

Figure 2: Screen shot of Argos showing various visualizations of analysis process. Two data sets are shown. The top middle window shows an image of the data set using orthogonal projections and bottom right shows a direct volume rendering of the raw data. The top right window shows a parallel coordinate view of the calculated attributes of found features. On the lower left a scatter plot of attribute data of the data sets is shown.

channels to colors or viewpoints cannot be changed. Picking of features remains possible as position information of features is stored. In this way the image can be used as a quick replacement for the directly calculated display of the data. Also the image can be combined with other spatial data saved during the batch processing just as can be done in the case of interactive inspection. For example the location of center of gravity of a feature can be indicated on the image.

To get a quick overview of a data set, an additional image attribute was added to the system which shows three orthogonal views of a data set. This view gives the same mapping possibilities as during interactive manipulation. An example is shown in the top middle window in figure 2. The first channel of the volume data is mapped to white and the outline of the selected features are mapped to red. Along with the image the physical dimensions of the depicted data set is shown. Using these values it is possible to show the locations of features and to allow picking of features in the image.

Interactive processing and batch processing can be combined during inspection of the data. The attribute data produced by the batch process can be loaded in the interactive environment and are treated in the same way as if the data was produced interactively. The images serve as a replacement for presentation of the real volume data. In some cases the images prepared during off line calculation do not offer enough information (e.g., when a certain slice in the data is of importance or when in the preset viewpoint of a feature of interest is obscured). In this case the user can repeat the analysis for a particular data set within the interactive environment recreating all intermediate data which then can be inspected using the available tools.

4.4 Visualization tools

The Argos system integrates quantitative analysis and visualization. Visualization is used to present extracted data as well as intermediate analysis results. This enables close inspection of the analysis process. The presentation of the data combines information visualization of the objects with attributes in a graph display (see figure 2 top right and bottom left), with direct visualization of volume data in a volume view (see figure 2 top middle and bottom right). Objects in the graph displays and the volume data are linked in several ways to improve the analysis process. The system operates such that when a data set is selected, the corresponding volume or image view is opened and when a feature is picked in one of both views it is highlighted in the other.

In the volume view the spatial aspects of the data can be studied (e.g., shape and location of spots and the volume data using volume rendering and slices). If available, the data is shown using mapping parameters which can be adapted interactively. If no data is available, the user can pick one of the generated visual summaries to be shown. In the graphing views, objects with their attributes can be inspected. There are separate views for the collection of data sets and for the collection of features found in the data sets. Information visualization is used to present and analyze the data. Objects can be selected to find attribute values of individual objects. Also the collection of objects can be graphed using the attribute values in various ways:

Histograms and box plots

Histograms and box plots are used to show the distribution of a single attribute value of a feature within a collection of data sets. The main use of histograms and box plots is presentation of final results in publications because they are a well known and easy to understand way to present scalar distributions (see for example figure 7 and 8).

Scatterplot

Scatterplots can be used to plot two attributes of objects in a graph. In the graph, regression lines and correlation values can be shown (see for example figure 5).

Parallel coordinates

Parallel coordinates [16] can be used to show the relative values of a large number of attributes of an object in a single plot.

Multidimensional scaling

Using multidimensional scaling [9], a collection of points in n-dimensional space is layed out in two dimensions such that n-dimensional distances between any two points is best preserved. This can be used to study the distribution of objects in the multidimensional attribute space for example to detect clusters in the collection.

The graphing and volume views on the data are linked in several ways. If a particular data set is selected in the graphing window the available volume data, either the volume data itself or images generated during batch processing, are shown in the volume view. Also all features found in the particular data set are highlighted in the feature graphing view. The other way around if a feature is selected the corresponding data set is highlighted as well.

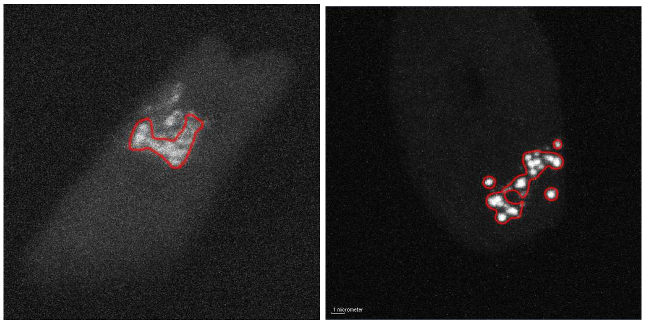


Figure 3: An example of the selection of the feature, i.e., the chromatin region. We used a Gaussian filter, followed by setting a threshold on the data. The sigma for the Gaussian and threshold were determined interactively for the selected set. The image represents the maximum intensity projection of the raw data. The outline in red shows the selected amplified region. The left image is from the control group, the right image shows a set from the full length HP1 targeted group.

5 Case study using Argos: Chromatin structure and gene control

Research to unravel gene control systems is a key focus in cell biology, presently receiving increasing international interest. Proper regulation of gene expression is crucial to maintain differentiated cell types in multicellular organisms [17]. Packaging of the eukaryotic genome into higher order chromatin structures is tightly related to gene expression [24, 25]. The nucleosome is the most basic level of the chromatin structure consisting of approximately 150 base pairs DNA wrapped around a core histone [22].

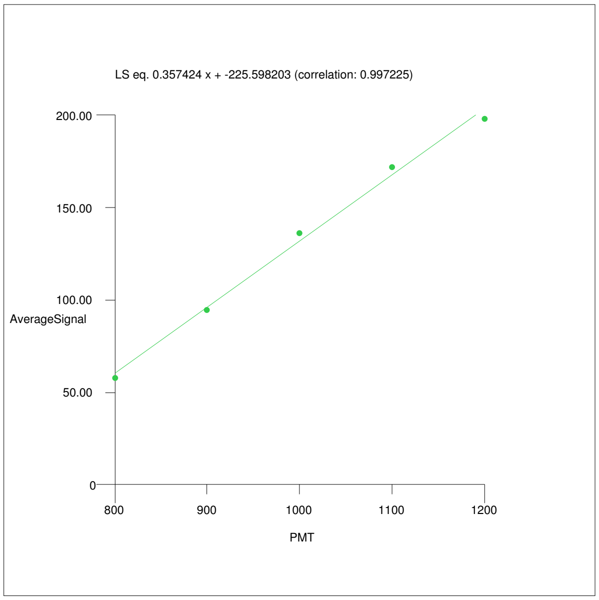
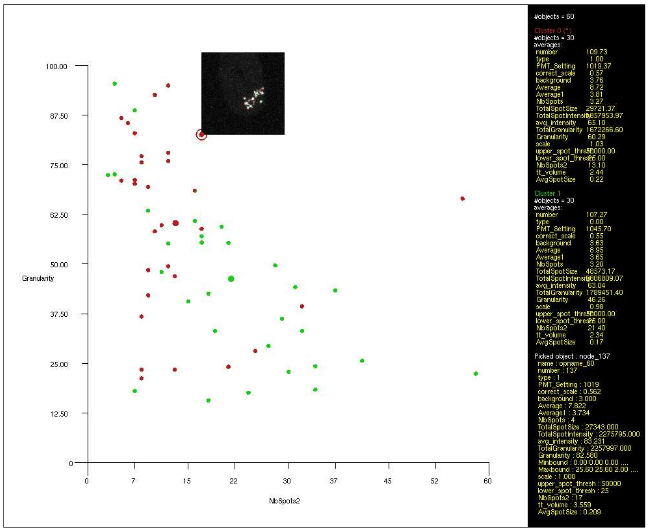


Figure 4: Derivation of a scaling parameter using Argos. The scatterplot shows the PMT-setting vs. average signal value and a linear least square fit.

Genome packaging phenomena act at different levels: (i) chromatin structural changes both at nucleosomal and higher order chromatin folding level and (ii) positioning of the genome within the 3D interphase cell nucleus at the nuclear level [15]. Little is known about in vivo mechanisms that establish and maintain functional genome organization, but posttranslational histone modifications and DNA methylation play a role in these processes. In the research team of PJV at the SILS, UvA, we aim to unravel the complex interacting systems of gene control using high resolution light microscopy in fixed and living cells as well as electron microscopy approaches in combination with sophisticated 3-D image analysis and processing tools [28, 7, 10, 27, 8]. We used the Argos system to analyze our data within the context of our research concerning chromatin structure and gene control. We focused on the effect of targeting a protein, Heterochromatin protein 1 (HP1) that is involved in gene regulation, to a defined chromosomal domain in the cell nucleus [26, 6]. We used a cell line containing a large 200 Mbp chromosome domain consisting of lac operator repeats. Such a large chromosomal domain can be visualized and followed in time at the light microscopy level by means of GFP tagged lac repressor binding to the lac operator repeats. We analyzed the effect of targeting both full length HP1 and a truncated HP1 version both as lac repressor-GFP-tagged fusion proteins to the amplified chromosome regions using targeting of lac repressor tagged GFP without HP1 as a control. Thirty cells of each test case were compared, i.e., control cells and cells in which either full length or truncated HP1 was targeted. Argos was used to quantitatively determine the effect of targeting full length or truncated HP1 on the chromatin structure at the chromosomal domain in these large data sets representing 3D CLSM images of the experiment.

The difference in intensity settings of the images was an initial problem for comparison of the data. The images were taken with different settings of the gain of the photo multiplier tube, to prevent clipping for intense spots occurring in some of the images. For



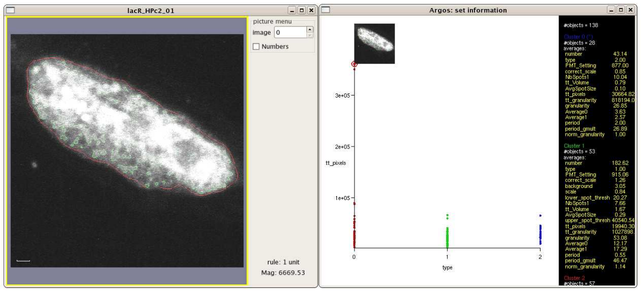
```
#objects = 60

Cluster 0 (*)
#objects = 30
averages:
 number              109.73
 type                  1.00
 PMT_Setting        1019.37
 correct_scale         0.57
 background            3.76
 Average               8.72
 Average1              3.81
 NbSpots               3.27
 TotalSpotSize     29721.37
 TotalSpotIntensity1857953.97
 avg_intensity        65.10
 TotalGranularity 1672266.60
 Granularity          60.29
 scale                 1.03
 upper_spot_thres50000.00
 lower_spot_thres25.00
 NbSpots2             13.10
 tt_volume             2.44
 AvgSpotSize           0.22

Cluster 1
#objects = 30
averages:
 number              107.27
 type                  0.00
 PMT_Setting        1045.70
 correct_scale         0.55
 background            3.63
 Average               8.95
 Average1              3.65
 NbSpots               3.20
 TotalSpotSize     48573.17
 TotalSpotIntensi2806809.07
 avg_intensity        63.04
 TotalGranularity 1789451.40
 Granularity          46.26
 scale                 0.98
 upper_spot_thres50000.00
 lower_spot_thres25.00
 NbSpots2             21.40
 tt_volume             2.34
 AvgSpotSize           0.17

Picked object : node_137
  name : opname_60
  number : 137
  type : 1
  PMT_Setting : 1019
  correct_scale : 0.562
  background : 3.000
  Average : 7.822
  Average1 : 3.734
  NbSpots : 4
  TotalSpotSize : 27343.000
  TotalSpotIntensity : 2275795.000
  avg_intensity : 83.231
  TotalGranularity : 2257997.000
  Granularity : 82.580
  Minbound : 0.00 0.00 0.00 ....
  Maxbound : 25.60 25.60 2.00 ....
  scale : 1.000
  upper_spot_thresh : 50000
  lower_spot_thresh : 25
  NbSpots2 : 17
  tt_volume : 3.559
  AvgSpotSize : 0.209
```

Figure 5: Scatterplot of set attributes. Graphing window shows information on data set attributes (granularity and number of distinct spots). Color shows the type of population, control population in green and full length HP1 targeted population in red. The two populations can be discriminated based on the shown attributes. A thumbnail of an image indicates the selected data set. On the right information on average values is displayed



```
#objects = 138

Cluster 0 (*)
#objects = 28
averages:
 number              43.14
 type                 2.00
 PMT_Setting        877.00
 correct_scale        0.85
 NbSpots1            10.04
 tt_Volume            0.79
 AvgSpotSize          0.10
 tt_pixels        30664.82
 tt_granularity   818194.0
 granularity         26.85
 Average0             3.63
 Average1             2.57
 period               2.00
 period_gmult        26.89
 norm_granularity     1.00

Cluster 1
#objects = 53
averages:
 number             182.62
 type                 1.00
 PMT_Setting        915.06
 correct_scale        1.26
 background           3.05
 scale                0.84
 lower_spot_thresh   20.27
 NbSpots1             7.66
 tt_Volume            1.67
 AvgSpotSize          0.29
 upper_spot_thresh 40540.54
 tt_pixels        19940.30
 tt_granularity   1027898.
 granularity         53.08
 Average0            12.17
 Average1            17.29
 period               0.55
 period_gmult        46.47
 norm_granularity     1.14

Cluster 2
#objects = 57
```

Figure 6: When the user picks a data set in the scatterplot (shown right) the visual summary of the picked data set is shown as a thumbnail. If needed the visual summary can be shown full size for further inspection as shown on the left. The y-axis of the scatterplot indicates the size of the selected area. In the shown set, the selection of the feature (chromatin region) is incorrect. Due to a high background level the whole nucleus is selected instead of the chromatin region. The color in the scatterplot indicates the type of population, i.e., control (red) full lenght HP1 targeted (green) and truncated HP1 targeted (blue).

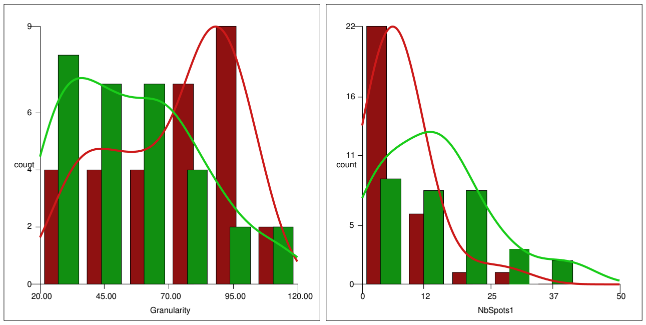


Figure 7: Histogram distribution of two attributes over the feature, left the granularity and right the number of distinct spots. Red bars represent the HP1 targeted population and green bars the control population. Histograms show that the distribution of these attributes can be discriminated fropm each other.

quantitative comparison of the images scaling had to be applied to the set intensities of the images such that the difference in PMT settings are corrected. To achieve this, a test specimen was imaged using a range of PMT-gain settings. In a simple setup the average signal intensity of the test specimen imaged with different PMT-gain settings was calculated. The output was used to produce a linear scaling function which mapped a PMT-setting to a scaling factor (see figure 4). This scaling factor was used as a parameter in a scaling operator to obtain comparable data sets in the setup.

In the setup phase five data sets of each type where used. Using this collection the operations and processing parameters of the analysis where determined. The first step was the extraction of the fluorescently labeled chromatin area from the data. This region is characterized by a higher intensity than the rest of the image. An example of the feature selection is shown in figure 3, the red outline represents the selected feature (the chromatin region). The chromatin region was selected by using a Gaussian filter, followed by setting a threshold on the data. The sigma for the Gaussian and threshold were determined interactively for the selected set. For the studied data collection, a value of sigma of 250 nm and a threshold of 30 percent of the maximum value gave the best result. The outcome of the selection was judged based on the visible selection of the chromatin area.

For the selected areas (features) a number of attributes were calculated:

- The volume of the feature, approximated by multiplying the number of voxels in the feature by the volume of a voxel.
- The average intensity gradient (granularity) of the feature. For each voxel the intensity gradient g_v is calculated by

  $$g_v = \sqrt{\Delta x^2 + \Delta y^2}.$$

  In the equation Δx and Δy denote the difference between intensity values of the voxel 4 voxels to left and right for the x direction and 4 voxels to the top and bottom for the y direction. The granularity is the average intensity gradient of the voxels in the feature.
- The local maximum intensity within the feature.
- The number local maximum values within the feature.
- The number of distinct spots within the feature. Distinct spots are determined by grouping EGFP-labeled voxels which share a face with another EGFP-labeled voxel.
- The average volume of the distinct spots within the feature.

The average value of the feature attributes are attributes of the data sets. Scatter plots revealed which of the calculated attributes can be used to distinguish the different populations of cells of the selected data sets. The granularity of the feature, the number of locally maximal values, and the size of spots in the selected region where different for the control versus HP1 targeted cells. Figure 5 shows the scatterplot of the granularity and the number of distinct spots. In the second phase, the feature extraction setup and analysis of the attributes of the selected feature where applied to the full set

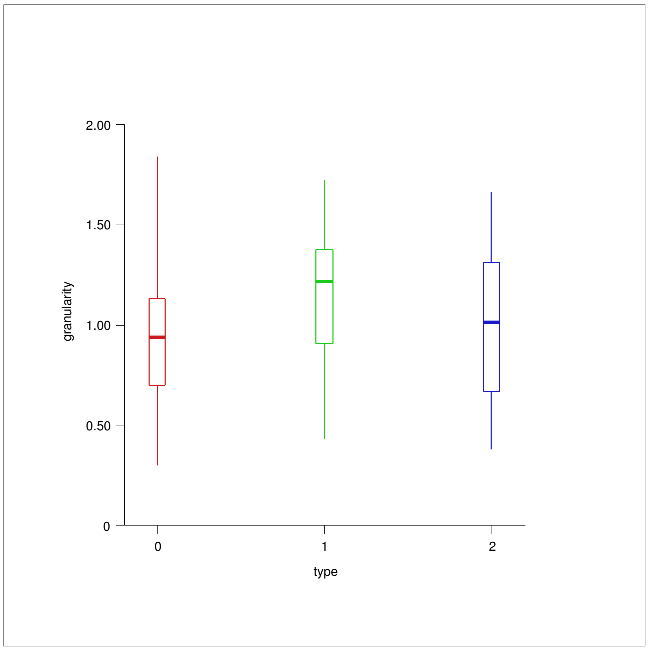


Figure 8: Boxplot representation of the granularity over the feature. Red, green and blue represent the control population, the full length HP1 targeted population, and the truncated HP1 targeted population, respectively. Statistical analysis using the Wilcoxon nonparametric test shows that both the full length HP1 targeted population and the truncated HP1 targeted population are significantly different from the control population.

of data sets. Inspection (see figure 6) of the visual summary images indicated that for two data sets the processing detection of the chromatin region failed. In these cases the selected feature clearly did not represent the fluorescently labeled signal i.e. the chromatin region. This incorrect selection is caused by high background levels in the image and/or high noise levels in the image. These sets with the incorrectly selected feature where excluded from the final result. Based on the distribution of at least three different attributes, i.e. the number of local maxima, the size of spots around the local maxima, and the granularity within the feature, the control and full length HP1 targeted cells could be discriminated from each other. Figure 7 shows the histogram distribution of values for the number of distinct spots (right) and granularity (left) for the control cells versus the HP1 targeted cells. The control sets are shown in green and the HP1 sets are shown in red. The histograms show that the distribution of both attributes can be discriminated in the two populations although there is overlap in the distribution of values. To test the quantitative significant difference between the data sets representing the cells in which full length HP1 was targeted versus the cells in which truncated HP1 was targeted, we used only one attribute, i.e., the granularity within the feature. The data are represented in a box plot (see figure 8). The second and third quartile of the observed values are within the box, the median value is shown by the thick horizontal line, the vertical small lines show the first and fourth quartile of the observed values. Statistical analysis using the Wilcoxon nonparametric test shows that both the full length HP1 targeted population and the truncated HP1 targeted population are significantly different from the control population ($p < 0.0001$).

6 Conclusions

In this paper we present a visualization approach for the analysis of large data collections. Tools are discussed to integrate batch processing and interactive processing. These tools are implemented in the Argos system. Data collections which are too large for interactive processing can be effectively analyzed with the Argos system. The main benefits of the Argos system are the combination of (i) an interactive setup of the batch process in which the the batch process can be tuned using one or a few data sets, (ii) visual summaries which are generated during the batch process and (iii) tools to combine the numerical analysis results with the visual summaries.

We applied the Argos system to the analysis of collections of confocal data sets in our research to unravel the changes in chromatin structure upon gene regulation. We visualized the chromatin structure making 3D image stacks with the CLSM. The Argos system allowed us to extract the feature, i.e. the chromatin region, from the data in the collections of data. Also, the Argos system allowed us to determine the significant difference between several data populations (control cells versus HP1 targeted cells) based on various attributes.

In biological experiments often large data collections are generated to overcome large biological diversity. From the present biological case study in which we used the Argos system, we conclude that the system is very helpful to extract relevant measures, to interactively test a part of the large data collection and to monitor the analysis procedure during processing.

6.1 Acknowledgments

We would like to thank Erik Manders for fruitful discussions on the analysis of the data. This work was partially carried out in the context of the Virtual Laboratory for e-Science project (www.vl-e.nl). This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).

References

[1] http://www.bitplane.com/.

[2] http://www.itk.org.

[3] http://www.kitware.com/products/volview.html.

[4] http://www.vaytek.com/voxblast.html.

[5] T. Biddlecomei, S. Fang, K. Dunn, and M. Tuceryan. Image guided interactive volume visualization for confocal microscopy data exploration. In *Proceedings 1998 SPIE International Symposium on Medical Imaging*, pages 130–140, 1998.

[6] M.C. Brink, Y. der Velden, W. de Leeuw, J Mateos-Langerak, A.S. Belmont, R. Van Driel, and Verschure P.J. Truncated hp-1 lacking a functional chromodomain induces heterochromatinization upon in vivo targeting. *Histochemistry and Cell Biology*, 125:53–61, 2006.

[7] D. Cmarko, P.J. Verschure, T.E. Martin, S. Krause, X.D. Fu, R. van Driel, and S. Fakan. Ultrastructural analysis of transcription and splicing in the cell nucleus after bromo-utp microinjection. *Molecular Biology of the Cell*, 10:211–223, 1999.

[8] D. Cmarko, P.J. Verschure, A.R. Otte, R. van Driel, and S. Fakan. Polycomb group gene silencing proteins are concentrated in the perichromatin compartment of the mammalian nucleus. *Journal of Cell Science*, 116:335–343, 2003.

[9] T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994.

[10] W.C. de Leeuw, R. van Liere, P.J. Verschure, A.E. Visser, E.M. Manders, and R. van Driel. Visualization of time dependent confocal microscopy data. In T. Ertl, B. Hamann, and A. Varshney, editors, *Proceedings IEEE Visualization 2000*, pages 473–476, Los Alamitos (CA), 2000. IEEE Computer Society Press.

[11] H.-C.Hege D.Stalling, M.Westerhoff. Amira - a highly interactive system for visual data analysis. In C.R. Johnson and C.D. Hansen, editors, *The Visualization Handbook*, pages 749–767. Elsevier, 2005.

[12] S. Fang, T. Biddlecome, and M. Tuceryan. Image-based transfer function design for data exploration in volume visualization. In D. Ebert, H. Hagen, and H. Rushmeier, editors, *Proceedings IEEE Visualization '98*, pages 319–326. IEEE Computer Society Press, 1998.

[13] R. Fernandez-Gonzalez, A. Jones, E. Garcia-Rodriguez, P.Y. Chen, A. Idica, M.H. Barcellos-Hoff, and C. Ortiz de Solorzano. A system for combined three-dimensional morphological and molecular analysis of thick tissue samples. *Microscopy Research and Technique*, 6(59):522–530, 2002.

[14] W. Gu, J. Vetter, and K. Schwan. An annotated bibliography of interactive program steering. *SIGPLAN Notices*, 29(9):140–148, 1994.

[15] P.J. Horn and C.L. Peterson. Molecular biology: Chromatin higher order folding: Wrapping up transcription. *Science*, 297:1824–1827, 2002.

[16] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1:69–91, 1985.

[17] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet*, 33:245–254, 2003.

[18] T. J. Jankun-Kelly and Kwan-Liu Ma. A spreadsheet interface for visualization exploration. In Thomas Ertl, Bernd Hamann, and Amitabh Varshney, editors, *Proceedings IEEE Visualization 2000*, pages 69–76. IEEE Computer Society, October 2000.

[19] T.J. Jankun-Kelly and Kwan-Liu Ma. Visualization exploration and encapsulation via a spreadsheet-like interface. *IEEE Transactions on Visualization and Computer Graphics*, 7(3):275–287, July/September 2001.

[20] A. Kaufman, R. Yagel, R. Bakalash, and I. Spector. Volume visualization in cell biology. In A. Kaufman, editor, *Proceedings IEEE Visualization'90*, pages 160–167. IEEE Computer Society Press, October 1990.

[21] D.W. Knowles, C. Ortiz de Solorzano, and S.J. Lockett. Analysis of the 3d spatial organization of cells and sub cellular structures in tissue. In D.L. Farkas and R.C. Leif, editors, *Proceedings of SPIE Vol. 3921*, pages 66–73, 2000.

[22] K. Luger, A.W. Mader, R.K. Richmond, D.F. Sargent, and T.J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389:251–260, 1997.

[23] G. Sakas, M.G. Vicker, and P.J. Plath. Visualization of laser confocal microscopy datasets. In R. Yagel and G.M. Nielson, editors, *Proceedings IEEE Visualization '96*, pages 375–380. IEEE Computer Society Press, 1996.

[24] R. van Driel, P.F. Fransz, and P.J. Verschure. The eukaryotic genome: a system regulated at different hierarchical levels. *Journal of Cell Science*, 116:4067–4075, 2003.

[25] P.J. Verschure. Positioning the genome within the nucleus. *Biol. Cell*, 96:569–577, 2004.

[26] P.J. Verschure, I. van der Kraan, W. de Leeuw, J. van der Vlag, A.E. Carpenter, A.S. Belmont, and R. van Driel. In vivo hp1 targeting causes large-scale chromatin condensation and enhanced histone lysine methylation. *Molecular and Cellular Biology*, 25:4552–4564, 2005.

[27] P.J. Verschure, I. van der Kraan, J.M. Enserink, M.J. Mone, E.M. Manders, and R. van Driel. Large-scale chromatin organization and the localization of proteins involved in gene expression in human cells. *J.Histochem. Cytochem*, 50:1303–1312, 2002.

[28] P.J. Verschure, I. van der Kraan, E.M. Manders, and R. van Driel. Spatial relationship between transcription sites and chromosome territories. *J. Cell Biology*, 147:13–24, 1999.

[29] R. Whitaker, D. Breen, K. Museth, and N. Soni. Segmentation of biological volume datasets using a level set framework. In A. Kaufman, editor, *Volume Graphics*, pages 249–263, Vienna, 2001. Springer.